Compression-Based Linguistic Complexity Metrics in Automatic Essay Scoring

Felipe Ribas Serras[1], **Igor Cataneo Silveira**[1], **Denis Deratani Mauá**[1], **Marcelo Finger**[1]

[1]Department of Computer Science
Institute of Mathematics, Statistics and Computer Science (IME)
University of São Paulo (USP) – São Paulo, SP – Brazil

{frserras, igorcs, ddm, mfinger}@ime.usp.br

Abstract. *Compression-based linguistic complexity metrics enable cross-linguistic comparison without prior annotation. Their sensitivity to variation across languages and Portuguese registers highlights their applicability in NLP tasks. This study investigates their use as readability proxies and complementary features in Automatic Essay Scoring. We analyze how these metrics capture variation in essay quality across traits, genres, and educational levels in Brazilian Portuguese. In addition, we evaluate their sensitivity to differences between human- and AI-generated essays. Our results suggest that complexity metrics are effective (i) in differentiating educational levels, (ii) in detecting whether they were written by humans and (iii) as predictors of essay quality.*

1. Introduction

Variation in structural linguistic complexity, the way in which linguistic manifestations employ strategies to distribute information across different linguistic subsystems, is a broad linguistic phenomenon that spans both the typological level, in terms of structural variation between languages [Croft 2002], and the sociolinguistic level, regarding how a single language adapts to different social contexts [Biber and Conrad 2009].

Structural complexity is an important parameter in various linguistic hypotheses and models, such as the equi-complexity hypothesis, the morphosyntactic trade-off hypothesis [Hockett 1958], and Multidimensional Analysis models of intralinguistic variation [Biber 1995]. It is also a major feature in the development and understanding of language processing technologies, such as linguistic simplification and elaboration models [Leal et al. 2026].

Grounded in information theory, compression-based linguistic complexity metrics [Juola 1998, Juola 2008, Ehret and Szmrecsanyi 2016] provide an automated approach to estimating structural linguistic complexity. By using off-the-shelf data compressors and random text perturbation, they measure the information conveyed at specific linguistic subsystems, such as syntax and morphology.

Because they operate directly on raw text rather than relying on annotation frameworks, these metrics are holistic and function comparably across diverse languages and linguistic scopes. Consequently, they offer a highly robust tool for analyzing transversal complexity phenomena, capturing evidence that emerges simultaneously across different scopes of linguistic variation, both intralinguistic and cross-linguistic. These metrics have demonstrated their descriptive power and sensitivity to different forms of variation, including typological variation across different languages and linguistic groups

[Juola 2008, Serras et al. 2024], and diachronic variation across historical variants of the same language [Ehret and Szmrecsanyi 2016].

In particular, this family of metrics has proven especially sensitive to register variation in both English [Ehret 2021] and Portuguese [Serras and Finger 2026]. Registers are varieties within a single language that emerge as a functional response to the context of use. Differences in register may vary in intensity, but they are pervasive throughout a text and are socially shared and mediated by a linguistic community, such as informal conversation, online forum discussions, and legal petitions [Biber and Conrad 2009].

One context in which the pervasive adaptation of language to a target register is instrumental is that of essay assessments, where candidates seek to demonstrate linguistic mastery by adhering to that register. Given the sensitivity of compression-based metrics to register variation, it is worth investigating how sensitive these metrics are to grade variation within this context, and whether they could serve as interpretable or supplementary features in automated essay scoring systems.

Automated Essay Scoring (AES) is an NLP task that aims to assign grades within a pre-established trait framework based on an essay's content [Page 1966]. These systems can serve as (i) support tools for graders, (ii) quality control mechanisms to ensure consistency in grading for institutions, and (iii) feedback proxies for student self-practice, especially for those with limited access to formal educational resources. In Brazilian Portuguese, these systems focus specially on essays required for the ENEM [Amorim and Veloso 2017, Marinho et al. 2021, Silveira et al. 2024, Rossman et al. 2026], the Brazilian National High School Exam, a large-scale and high-stakes exam that serves as the national standardized test for university admission, which requires the production of an argumentative essay on a different prompt every year.

The advent of Transformer-based Large Language Models has brought a significant increase in performance to AES systems [Boquio and Naval 2024]. In general, but especially for Portuguese, the state-of-the-art results in AES are currently those obtained by this family of models [Mello et al. 2024, Barbosa et al. 2025]. In this context, explainability is one of the greatest challenges for those systems, which function as black boxes. This makes it difficult to understand both the means by which the models achieve this performance and the rationale behind the scores they assign.

Readability metrics have being widely used in the context of AES for Portuguese, as interpretable features for models [Marinho et al. 2022, Silveira et al. 2025b, Chalegre et al. 2026] and also as means of distinguishing different essay sub-registers [Cavalcanti et al. 2025]. However, the inclusion of new features in these analyzes can consistently offer a fresh perspective on how AES models operate and the differences and peculiarities of the diverse range of essay sub-registers to which they are applicable.

To this end, we seek to integrate these two areas of investigation in this article by examining the sensitivity of compression-based linguistic complexity metrics to grade variation in AES datasets across different educational levels and essay sub-registers. By doing so, we intend to better understand the sensitivity and robustness of these metrics, generating a deeper understanding of their behavior in Portuguese and ascertaining their usability as AES features for both model training and analysis.

Inspired by [Serras and Finger 2026], we conducted a statistical analysis of the

sensitivity of these metrics, both to scores and to differences between sub-registers of essays, using three AES datasets in Brazilian Portuguese: the Brazilian Portuguese Narrative Essays Dataset [Mello et al. 2024], AES-ENEM [Silveira et al. 2024], and Diplomatrix-BR [Cavalcanti et al. 2025]. These datasets cover narrative and argumentative essays produced by children, adolescents, adults, and language models. Through this study, we aimed to answer the following research questions:

1. How well do compression-based complexity metrics predict the scores assigned to essays across different sub-registers, educational levels, and traits?
2. How well do these metrics distinguish among different essay sub-registers?
3. How well do these metrics distinguish between human-written and LLM-generated essays?
4. What do these results reveal about structural linguistic variation across essays and the sensitivity of this specific family of metrics?
5. Could these metrics be incorporated into AES systems as features, or aid in their interpretation?

Overall, we conclude that these metrics are quite effective at distinguishing among different essay sub-registers and between human-written and synthetic essays. In many cases, they serve as statistically significant predictors of the assigned scores, particularly where there is greater structural variation and where adherence to linguistic structure is assessed more directly, such as in essays written by children who are still mastering these structures. Nevertheless, they generally act as weak to moderate score predictors; while they can be valuable tools for interpreting these models, they should be used in conjunction with other metrics to provide a more accurate picture of their behavior.

2. Related Work

In this section, we review related work, categorizing it into the two main areas of research covered by this study: linguistic complexity metrics and AES.

2.1. Linguistic Complexity Metrics

The modern approach to structural linguistic complexity as a parameter stems from Hockett's hypotheses [Hockett 1958]. [Nichols 1998] and [McWhorter 2001] are examples of the earliest proposals to characterize this form of complexity as a quantitative parameter, involving subjective interpretation and manual annotation by experts.

Methods for automatically computing complexity first emerged through readability metrics in the field of language acquisition [Leal et al. 2026]. Tools such as Coh-Metrix [McNamara et al. 2014] for English, and NILC-Metrix [Leal et al. 2024] and iRead4Skills [Monteiro et al. 2025, Aissa et al. 2025] for Portuguese represent the culmination of this approach, enabling the automated extraction of various complexity and readability features from text. Although powerful, these tools rely on annotation systems that vary across languages, limiting their applicability in cross-linguistic parameterization.

Compression-based Complexity metrics have emerged as alternative text-based approaches that are grounded in information theory and language-agnostic, enabling the comparability of results across different languages and linguistic scopes.

Proposed in [Juola 1998, Juola 2008] and refined in [Ehret and Szmrecsanyi 2016], they were validated against cross-linguistic [Juola 2008, Serras et al. 2024], diachronic [Ehret and Szmrecsanyi 2016], and synchronic intralinguistic variation in English [Ehret 2021] and Portuguese [Serras and Finger 2026]. The latter studies explored variation at the register level, where broader linguistic descriptions from Multidimensional Analysis Models point to a strong dependence on structural complexity and informational density across various languages [Biber and Conrad 2009, Biber 1995], including Portuguese [Sardinha et al. 2014, Serras et al. 2026].

More recently, readability and complexity features have also been explored in the characterization and detection of synthetic texts generated by LLMs. [Lobo and Martins 2026], for example, examines the differences between Portuguese scientific abstracts with and without LLM influence, using NILC-Metrix features. [Cavalcanti et al. 2025] presents Diplomatrix-BR, an essay dataset comprising both human-authored texts and essays generated by various LLMs in response to the same prompts; furthermore, using NILC-Metrix features, the authors analyze the differences between these groups. Readability features were also used by [Locatelli et al. 2024] and [Liu et al. 2023] to compare synthetic and natural argumentative essays in Portuguese and English, respectively.

2.2. Automatic Essay Scoring

Although early AES systems relied on a feature-based approach combined with linear regression [Page 1966], subsequent NLP technologies improved model generalization at the expense of interpretability. However, in this domain, explicability is critical, and it is often worthwhile to sacrifice predictive performance in favor of interpretability. For this reason, modern approaches seek to balance both objectives.

For Portuguese, several sets of interpretable features have been used to this end. Examples are: the words in the essay [Bazelato and Amorim 2013]; a combination of domain-relevant features with features adapted from English [Amorim and Veloso 2017]; n-grams, and POS-tagging n-grams [Fonseca et al. 2018]; and ENEM-trait-specific feature sets [Marinho et al. 2022]. Recently, NILC-Metrix has been widely employed as it offers a comprehensive suite of features within a single framework. It has been used (i) to calculate proxies for what Transformer-based models may be learning [Silveira et al. 2025a], (ii) as baseline performance when combined with a linear regressor and random forest [Barbosa et al. 2025], and (iii) as an alternative input representation for neural networks [Chalegre et al. 2026].

Other works aim to combine Transformers and feature explainability. One approach combines them by adding the Transformer-assigned score to a pool of interpretable features [Liu et al. 2019, Silveira et al. 2025b], or by concatenating the features with dense representations [Matos and Feltrim 2026]. Alternatively, another approach employs Transformers to predict specific essay characteristics, subsequently combining them through (interpretable) logic [Silveira and Mauá 2026].

3. Methodology

In this section, we describe our complexity metrics (Section 3.1), the AES datasets that we examined (Section 3.2), the statistical methods used to assert the correlation between

the metrics and the grades in each dataset (Section 3.3), and the methods used to attest the sensitivity of the metrics to differentiate between essay sub-registers (Section 3.4). The code, plots, and results from all described procedures are available in our repository[1].

3.1. Compression-based Complexity Metrics

Compression-based complexity metrics reduce language to its communicative function of transmitting information. In this view, the complexity of a text T is defined as the amount of information it conveys or the shortest representation of the text from which it can be reconstructed, known as Kolmogorov complexity. The Kolmogorov complexity of T is generally uncomputable but can be approximated using data compression algorithms, such as `gzip`, by generating a compressed version $K(T)$ of the original text and measuring its size in bits, represented by $|K(T)|$ [Grünwald and Vitányi 2008].

To assess the amount of information conveyed by a specific linguistic subsystem μ, such as syntax s or morphology m, these metrics apply random perturbations to the text, designed to primarily affect the patterns of the target subsystem ($T \rightarrow T^{\mu}$), and measure the effect on the amount of information transmitted ($|K(T^{\mu})|$). The syntactic complexity C_s and morphological complexity C_m metrics are obtained by comparing this perturbed value to the amount of information in the original text, following Equations 1 and 2. In the version of these metrics used in this study, the perturbation applied to the texts consists of randomly deleting 10% of the units within the target subsystem: characters for morphology and entire words for syntax.

$$C_s(T) = \frac{|K(T^s)|}{|K(T)|} \quad (1) \qquad C_m(T) = -\frac{|K(T^m)|}{|K(T)|} \quad (2) \qquad C^*_m(T) = \frac{|K(T)|}{|K(T^{m*})|} \quad (3)$$

Conceptually, C_s measures the importance, in terms of information transmission, of a specific word appearing in a specific position in the text; it therefore estimates the obligatory level of syntactic elements: the more restricted and closed the structure is regarding the presence of its elements, the higher C_s will be. C_m, on the other hand, is a measure of lexical-morphological diversity at the informational level. The more diverse the lexicon and the lexical construction mechanisms in the sample, the higher we expect C_m to be.

An alternative measure of morphological complexity is C^*_m, presented in [Juola 1998]. For this measure, the perturbation applied to morphology is more significant, replacing all words with unique numerical indices. Thus, the perturbed version of the text, T^{m*}, is stripped of all information about the content of the words themselves, retaining only information about their positions. The resulting C^*_m complexity, computed using Equation 3, measures the amount of raw information contained in word representations. In this sense, it is a much coarser measure than C_m, as it seeks to capture the overall morphological information, rather than the subtleties of the sample's morphological system.

For our experiments, we used [Serras et al. 2024]'s implementation of these metrics[2], which computes C_s, C_m, and C^*_m in Python using two distinct compressors: `gzip`, a general-purpose data compressor, and `bzip2`, a compressor optimized for text data.

[1] `https://github.com/frserras/essay-complexity-pt`

[2] `https://github.com/frserras/lang_complexity` (`version 1.0.0`)

3.2. AES Datasets

In this study, we used three datasets:

- The **Brazilian Portuguese Narrative Essays Dataset** [Mello et al. 2024] which comprises 1233 narrative essays written by students from 5th to 9th grade in Brazil on two different prompts. These essays are graded in an ordinal scale containing five classes, according to four different traits: `cohesion`, `formal register`, `narrative rhetorical structure` and `thematic coherence`.
- The **AES-ENEM Dataset** is composed by argumentative essays written for mock ENEM exams [Silveira et al. 2024]. Here essays are also graded in an ordinal scale containing six classes, in five different traits: `(C1) proper usage of Portuguese`, `(C2) thematic coherence`, `(C3) argumentation`, `(C4) coherence` and `(C5) conclusion strength`. Importantly, we removed the lowest class, as it does not reflect text quality but rather the author not adhering to the prompt. This dataset includes grades by two different graders and the original web scrapped grades.
- The **Diplomatrix-BR dataset** [Cavalcanti et al. 2025] is made of argumentative essays written for diplomat civil service tests. This last dataset differs from the previous ones in two ways: a) it is scored on a continuous scale; b) it contains both synthetic and human-written essays.

3.3. Grade-Complexity Correlation Analysis

To assess the predictive power of complexity metrics for essay scores, we computed each metric variant listed in Section 3.1 ten times per essay and averaged the results. We then measured the correlation between each metric and the scores for each trait within each dataset separately.

We chose to categorize the Diplomatrix continuous scores into quintiles, transforming them into an ordinal categorical variable, consistent with the scores in the other two datasets. This allowed us to apply the same correlation test across all datasets. We used Kendall's rank correlation test [Hollander et al. 2014], specifically designed for ordinal categorical variables[Brideau 2025, Laerd Statistics 2018]. Of the three correlation coefficients frequently used for this test, we selected $\tau - b$, which is recommended for scenarios with ties, such as ours [Kendall 1945].

The test is non-parametric [Hollander et al. 2014] and relies on only two assumptions: (i) both variables must be ordinal, which holds true for both complexity metrics and scores; and (ii) there must be an approximately monotonic relationship between the target variables. This latter assumption is not strict; however, when the relationship deviates from monotonicity, there is an increased risk of false negatives [Laerd Statistics 2018]. We examined the relationships across all metric-score pairs: most were approximately monotonic, but we isolated the non-monotonic pairs for a more cautious interpretation.

Since multiple tests were performed simultaneously, we applied the Benjamini-Hochberg False Discovery Rate correction to the p-values to reduce the risk of false positives resulting from multiple testing [Benjamini and Hochberg 1995]. For both the test and the p-value correction, we used the implementations from the `scipy`[3] package.

[3]https://scipy.org/

3.4. Essay Sub-register Sensitivity Analysis

To assess the ability of the metrics to differentiate the essay sub-registers, we used the analysis methodology outlined in [Serras and Finger 2026]: (i) we segmented the essays from all datasets into sentences using the Portuguese sentence tokenizer from the spaCy package[4]; (ii) we randomly selected 2200 sentences from each sub-register to compile the sentence dataset; (iii) we sampled from this dataset 250 sentences per sub-register, combining them in random order to form a pseudo-document, repeating this process 250 times for each sub-register to generate 250 pseudo-documents; (iv) we computed the complexity metrics for each pseudo-document; and (v) we used these values to perform an Analysis of Variance (ANOVA) test to determine if the complexity varies significantly across the sub-registers. In contrast to [Serras and Finger 2026], we employed Welch's ANOVA[5], which does not assume homogeneity of variances across groups, as our exploratory analyses revealed that the sub-registers exhibit markedly distinct width distributions across several metrics [Welch 1951, Liu 2015].

The test computes the F-statistic, which evaluates the ratio of between-group variability to within-group variability. A higher F-value indicates greater separability of the sub-registers for a given complexity metric. Because the F-statistic can be sensitive to sample size, we also calculated the effect size η^2. An η^2 value approaching 1 indicates highly separable sub-registers.

The assumptions of Welch's ANOVA are: (i) independence of observations, which was ensured by the experimental design, in which each data point corresponds to a pseudo-document with an approximate 10% sentence overlap; and (ii) normality of residuals, which we verified by analyzing Quantile-Quantile (Q-Q) plots comparing the observed residuals for each metric against a theoretical normal distribution [Welch 1951, Liu 2015].

We also investigated the morphosyntactic trade-off across essay sub-registers by (i) examining scatter plots of complexity metrics and (ii) assessing the correlation coefficients between these metrics. The morphosyntactic trade-off is a structural pattern wherein higher syntactic complexity tends to co-occur with lower morphological complexity, and vice versa. Compression-based metrics are particularly sensitive to this phenomenon, both at the typological level [Juola 2008, Serras et al. 2024] and across broader linguistic registers in Portuguese [Serras and Finger 2026]. This study allows us to verify whether this pattern persists at this more granular level of variation.

4. Results

The results are divided into two parts: Grade-Complexity Correlation Analysis (Section 4.1) and Essay Sub-register Sensitivity Analysis (Section 4.2).

4.1. Grade-Complexity Correlation Analysis

Figures 1a and 1b present the correlation matrices between the metrics and the score traits for the Narrative dataset, and the AES-ENEM Dataset[6] respectively. Cells marked

[4] `https://spacy.io/`

[5] We utilized the `statsmodels` library (`https://www.statsmodels.org/stable/index.html`)

[6] For brevity, we present results for "Grader A". Full results are in our repository

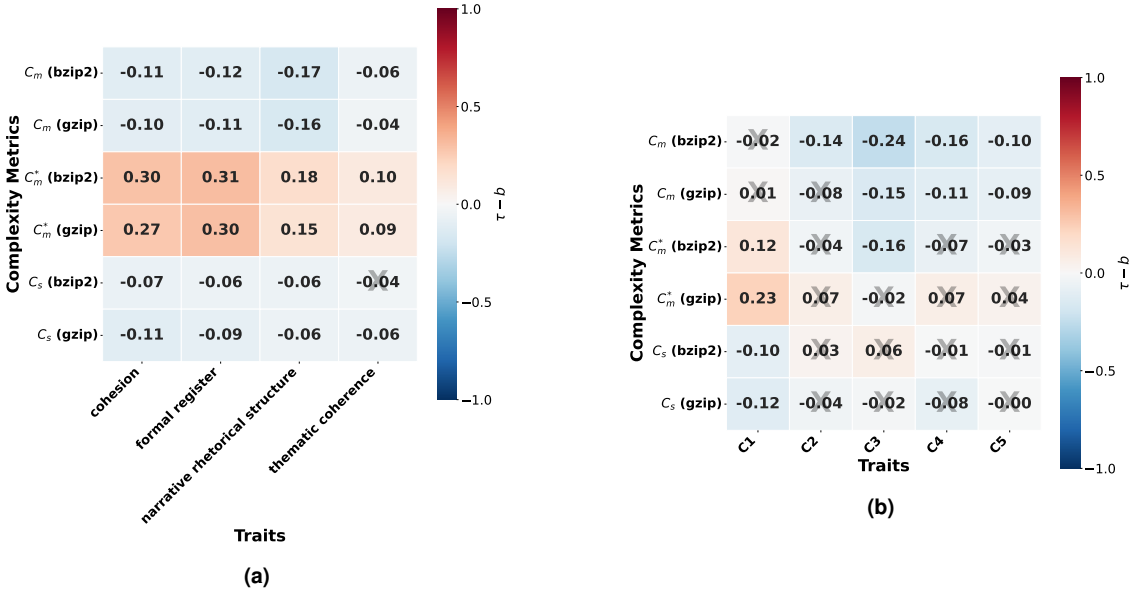

Figure 1. $\tau - b$ **correlation coefficients for (a) the Narrative Dataset and (b) the AES-ENEM Dataset. Cells marked with an X have non-significant associated p-values.**

with 'X" indicate statistically non-significant p-values and are therefore excluded from consideration. In several cases, we observed statistically significant correlation coefficients, indicating that the complexity metrics studied are meaningful predictors of scores. However, these correlations indicate low to moderate predictability, suggesting that these metrics should be used supplementarily rather than independently.

The higher the educational level of the essays, the weaker and less statistically significant the correlations tend to be. Notably, the metrics are not statistically significant predictors for differentiating the quintiles of the Diplomatrix scores, whereas they are significant predictors for all traits of the narrative dataset. These results can be explained by two factors: (i) at more advanced educational levels, candidates are more likely to demonstrate mastery of the desired linguistic structures, producing more cohesive essays with less structural variation; and (ii) at less advanced levels, structural adherence becomes more critical for grading due to substantial variation among candidates. Consequently, although adherence to register is an evaluative criterion at higher educational levels, it is seldom a differentiating factor because the vast majority of candidates achieve high adherence. Conversely, the narrative dataset exhibits notable variation; a qualitative sampling reveals some narrative texts that adhere closely to the expected register, while others present clearly non-adherent structures, especially polysyndetic sentences characteristic of children's writing.

Among all metrics, C_m^*, which represents raw morphological complexity, is the strongest predictor of grades. This metric correlates with text length and word length, indicating that, as expected, longer texts or those with greater lexical complexity or diversity tend to achieve higher evaluation scores. The traits most strongly correlated with this metric are `formal register` and `cohesion` in the Narrative Dataset, and `(C1) proper usage of Portuguese` in the AES-ENEM dataset. The studied structural complexity metrics, known to be sensitive to register variation, align precisely with the

traits assessing adherence to formal register and structure, as expected.

The other metrics exhibit weaker or inverse correlations with the scores, indicating that, at least within these samples, these subtle dimensions of complexity are negative predictors and less prominent than raw morphological complexity. We hypothesize that the direction of these correlations stems from C_m capturing off-topic lexico-morphological diversity and C_s reflecting excessive syntactic rigidity, which negatively impacts the scores. Remarkably, the part of the ENEM dataset that was graded by experts demonstrated higher inter-annotator agreement in terms of correlation with the complexity metrics than the part that was web-scraped, underscoring the consistency and quality of those annotations.

4.2. Essay Sub-register Sensitivity Analysis

Table 1 presents the ANOVA results evaluating the metrics' ability to differentiate writing sub-registers. The large F-statistics demonstrate the strong capacity of these metrics to differentiate registers, even at this higher level of specificity. The large η^2 values reinforce these findings, indicating that they are not an artifact of sample size. Overall, the discriminatory power of the morphological metrics C_m and C_m^* is greater than that of the syntactic metrics C_s, consistent with the observations of [Serras and Finger 2026].

Metric	F	P-Value	η^2
C_m (bzip2)	11344	< 0.001	0.98
C_m (gzip)	11376	< 0.001	0.98
C_m^* (gzip)	10658	< 0.001	0.97
C_m^* (bzip2)	3190	< 0.001	0.93
C_s (gzip)	1334	< 0.001	0.81
C_s (bzip2)	805	< 0.001	0.73

Table 1. ANOVA results between all essay sub-registers

Metric	F	P-Value	η^2
C_m (gzip)	7193	< 0.001	0.94
C_m (bzip2)	3766	< 0.001	0.88
C_m^* (gzip)	2951	< 0.001	0.86
C_s (gzip)	2167	< 0.001	0.81
C_s (bzip2)	1042	< 0.001	0.68
C_m^* (bzip2)	646	< 0.001	0.56

Table 2. ANOVA results between human-authored and LLM-generated essays

To distinguish between LLM-generated and human-authored texts, we conducted a separate ANOVA on the Diplomatrix data. The results, presented in Table 2, further demonstrate the robust capacity of the studied metrics to differentiate these two categories.

When evaluating the trade-off between syntactic and morphological complexity metrics, we excluded the narrative dataset. This subset exhibited significantly different complexity values, likely reflecting the characteristics of children's writing, which allowed for greater sensitivity in the previous analysis. Figure 2 illustrates the trade-off between the remaining sub-registers using the `gzip`-based metric variants. Despite the limited number of data points, the overall trade-off remains evident. Specifically, Figure 2b demonstrates a clear distinction between human-authored essays and those produced by LLMs.

The results suggest that LLM-generated texts are shorter, exhibit reduced lexical variety, and employ more formulaic and rigid syntactic structures than human-authored texts, consistent with the observations of [Liu et al. 2023, Locatelli et al. 2024]. This aligns with the hyper-structuring of these synthetic essays, which often organize the con-

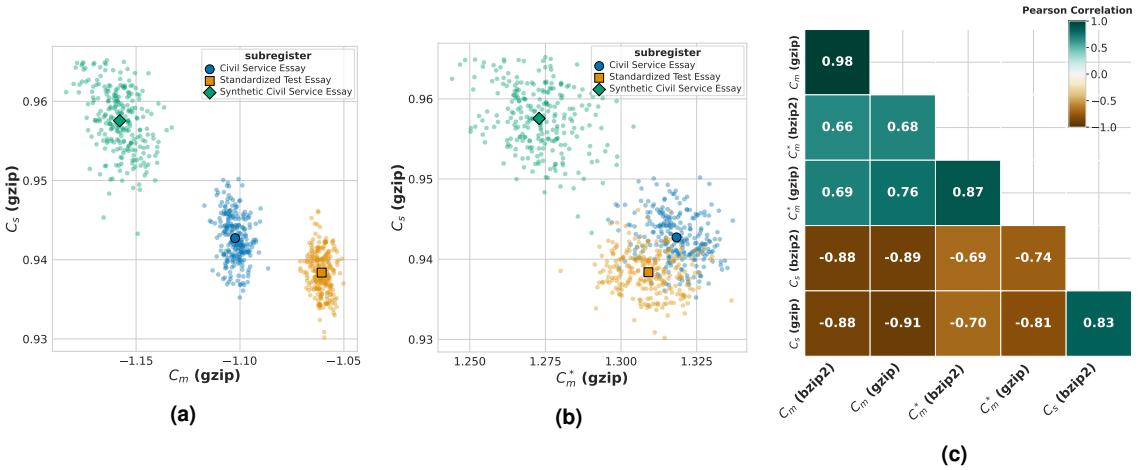

Figure 2. Trade-off plots between Syntactic and Morphological Complexity Metrics using `gzip` (a, b) and Correlation matrix between complexity metrics across essay sub-registers (c).

tent into sections and copy entire parts of the reference text, a behavior rarely observed in human writers.

Conversely, the ENEM Dataset essays exhibit less rigid syntax and greater lexico-morphological diversity than those produced for the diplomatic exam, in accordance with the observations in [Cavalcanti et al. 2025] about diversity and ambiguity. This likely reflects the broader thematic scope of ENEM essays, which cover more comprehensive topics; furthermore, their wider variance in register adherence results in lower syntactic complexity.

The correlations among the various metrics across the sub-registers, illustrated in Figure 2c, support these results, demonstrating an inverse correlation between syntactic and morphological metrics, positive correlations among metrics within the same subsystem, and strongly positive correlations between variants of the same metric using different compressors. This aligns with previous findings suggesting that, at least when using `gzip` and `bzip2`, the effect of variation in data complexity outweighs the variation introduced by the chosen compressor.

5. Conclusions

In this study, we evaluated the sensitivity of compression-based structural complexity metrics to variation in essay sub-registers, as well as their predictive power for scores across different educational levels and traits. Overall, the metrics serve as statistically significant predictors of grades, especially when adherence to the target register structure is a key grading criterion. However, their predictive power tends to be weak to moderate, indicating that they should be used as supplementary features. Furthermore, these metrics effectively differentiate between sub-registers as well as between LLM-generated and human-authored texts, exhibiting a morphosyntactic trade-off and an emphasis on morphological differentiation, consistent with previous findings.

Limitations

The limitations of this study and avenues for future work include (i) small sample sizes and confounded variables, such as the overlap between narrative and children's texts; (ii) the need for more diverse, multilingual datasets to leverage language-agnostic metrics; (iii) the potential of combining explainable metrics with qualitative analysis; (iv) unexplored options in the Diplomatrix dataset, such as alternative continuous-to-categorical conversion methods and model-specific segmentation of synthetic essays; (v) sub-optimal essay lengths for compression-based metrics, although exploring longer texts remains a future path; (vi) a scope limitation to robust statistical exploration without actual model training on the studied metrics, which is a natural next step; and (vii) the lack of direct quantification of the influence between C_m^* and text length, despite qualitative indications.

Ethics Statement

For this study, we utilized publicly available, open-source datasets and code in accordance with their respective usage licenses. Across all datasets, the essays were anonymized, with the exception of Diplomatrix-BR, where author data is public; however, even in this instance, we did not incorporate author-identifying information into any aspect of our research.

Generative AI usage

The authors used generative AI tools (Gemini 3 Pro) to assist with writing (orthographic correction, paraphrasing and language refinement) and code development. All AI-generated suggestions were verified by the authors.

Aknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the University of São Paulo, by the São Paulo Research Foundation (FAPESP) (grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) (grants 2023/00488-5 (SPIRA-BM), 2024/19144-7 (SPRINT)); and the National Council for Scientific and Technological Development (CNPq) (grant PQ 302963/2022-7). Felipe Ribas Serras was supported by IBM through a scholarship managed by FUSP (Support Foundation for the University of São Paulo) (Project 3541) and by a PPI-SOFTEX scholarship managed by FUSP (Project 3970). Denis Deratani Mauá was partly supported by FAPESP (grant 2022/02937-9) and CNPq (grant 305136/2022-4). Igor Silveira was supported by FAPESP (grant 2026/15768-1). This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

References

Aissa, W., Amaro, R., Antunes, D., Bañeras-Roux, T., Baptista, J., Catala, A., Correia, L., François, T., Garcia, M., Izquierdo-Álvarez, M., Mamede, N., Martins, V., Neves, M., Ribeiro, E., Rey, S. R., and Vanzeveren, E. (2025). The iRead4Skills intelligent complexity analyzer. In Habernal, I., Schulam, P., and Tiedemann, J., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 73–84, Suzhou, China. Association for Computational Linguistics.

Amorim, E. C. F. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102. Association for Computational Linguistics.

Barbosa, A., Silveira, I. C., and Mauá, D. D. (2025). An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Bazelato, B. S. and Amorim, E. C. F. (2013). A bayesian classifier to automatic correction of portuguese essays. In *Conferência Internacional sobre Informática na Educação (TISE)*, volume 18, pages 779–782.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, 1st edition.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press, 1st edition.

Boquio, E. N. V. and Naval, Jr., P. C. (2024). Beyond canonical fine-tuning: Leveraging hybrid multi-layer pooled representations of BERT for automated essay scoring. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2285–2295, Torino, Italia. ELRA and ICCL.

Brideau, L. (2025). Correlation: Pearson, spearman, and kendall's tau. UVA Library StatLab, University of Virginia.

Cavalcanti, R., Casini, G., Assis, G., Real, L., Vianna, D., Mann, P., and Paes, A. (2025). Diplomatrix-br: Um corpus paralelo de redaçoes de autoria humana e de llms no concurso de diplomacia brasileira. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 192–205. SBC.

Chalegre, P. C., Machado, V. d. R., and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Uma análise comparativa entre engenharia de características e transformers. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 738–748, Salvador, Brazil. Association for Computational Linguistics.

Croft, W. (2002). *Typology and Universals*. Cambridge University Press, 2 edition.

Ehret, K. (2021). An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics and Linguistic Theory*, 17(2):383–410.

Ehret, K. and Szmrecsanyi, B. (2016). *An information-theoretic approach to assess linguistic complexity*, pages 71–94. De Gruyter, Berlin, Boston.

Fonseca, E. R., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *Computational Processing of the Portuguese Language. PROPOR 2018.*, pages 170–179.

Grünwald, P. D. and Vitányi, P. M. (2008). Algorithmic information theory. In Adriaans, P. and van Benthem, J., editors, *Philosophy of Information*, Handbook of the Philosophy of Science, pages 281–317. North-Holland, Amsterdam.

Hockett, C. (1958). *A Course in Modern Linguistics*. A Course in Modern Linguistics. Macmillan.

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., Hoboken, New Jersey, 3rd edition.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Juola, P. (2008). *Assessing linguistic complexity*, pages 89–108. John Benjamins Publishing Company.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Laerd Statistics (2018). Kendall's tau-b using spss statistics - a how-to statistical guide.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: Assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Leal, S. E., Serras, F. R., Finger, M., and Aluísio, S. M. (2026). Complexidade textual e suas tarefas relacionadas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3, book chapter 6. BPLN, 4 edition.

Liu, H. (2015). Comparing Welch's ANOVA, a Kruskal-Wallis test and traditional ANOVA in case of heterogeneity of variance. Master's thesis, Virginia Commonwealth University, Richmond, VA. VCU Theses and Dissertations.

Liu, J., Xu, Y., and Zhu, Y. (2019). Automated essay scoring based on two-stage learning.

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., and Hu, H. (2023). Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models.

Lobo, T. R. and Martins, C. A. (2026). Evolução de padrões linguísticos na escrita científica em português: Uma análise com NILC-metrix. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of*

the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1, pages 1062–1067, Salvador, Brazil. Association for Computational Linguistics.

Locatelli, M. S., Miranda, M. P., Da Silva Costa, I. J., Prates, M. T., Thomé, V., Zaparoli Monteiro, M., Lacerda, T., Pagano, A., Rios Neto, E., Meira Jr., W., and Almeida, V. (2024). Examining the behavior of llm architectures within the framework of standardized national exams in brazil. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:879–890.

Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2021). Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64. Sociedade Brasileira de Computação.

Marinho, J. C., Cordeiro, F., Anchiêta, R. T., and Moura, R. S. (2022). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC.

Matos, G. G. d. and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Um estudo empírico sobre representações linguísticas e contextuais. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 488–497, Salvador, Brazil. Association for Computational Linguistics.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, 1st edition.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typology*, 5(2-3).

Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (2024). Propor'24 competition on automatic essay scoring of portuguese narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 1–5.

Monteiro, R., Correia, S., Amaro, R., Moutinho, M., Barbosa, S., and Reis, M. L. (2025). Níveis e descritores de complexidade textual para adultos de baixa literacia: um referencial do projeto iread4skills. *Revista da Associação Portuguesa de Linguística*, (13):193–222.

Nichols, J. (1998). *Linguistic Diversity in Space and Time*. University of Chicago Press.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, pages 238–243.

Rossman, L. N., Silveira, I. C., and Mauá, D. D. (2026). Evaluating automated scoring models on official ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 161–171, Salvador, Brazil. Association for Computational Linguistics.

Sardinha, T. B., Kauffmann, C., and Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in Brazilian Portuguese. *Corpora*, 9(2):239–271.

Serras, F., Carpi, M., Branco, M., and Finger, M. (2024). Analysing and validating language complexity metrics across South American indigenous languages. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., and Oseki, Y., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 152–165, Bangkok, Thailand. Association for Computational Linguistics.

Serras, F. R., Carpi, M. D. M., Sturzeneker, M. L., Palma, M. F., Costa, A. S., Monte, V. M. D., Namiuti, C., Crespo, M. C. R. M., Paixão De Sousa, M. C., and Finger, M. (2026). Análise e Classificação Automática de Domínios Discursivos no Português do Brasil. *Linguamática*, 17(2):131–171.

Serras, F. R. and Finger, M. (2026). Compression-based language complexity under register variation in Portuguese. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 808–818, Salvador, Brazil. Association for Computational Linguistics.

Silveira, I. C., Barbosa, A., da Costa, D. S. L., and Mauá, D. D. (2025a). Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 228–237.

Silveira, I. C. and Mauá, D. D. (2026). Neuro-symbolic approaches for rubric-based automatic essay evaluation of ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 790–799, Salvador, Brazil. Association for Computational Linguistics.

Silveira, I. C., Ribeiro, E., Mamede, N., and Baptista, J. (2025b). Aprendizado por transferência para correçao automática de redaçao. *Linguamática*, 17(2):99–116.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336.